UNITED STATES DEPARTMENT OF COMMERCE
Economics and Statistics Administration
U.S. Census Bureau
Washington, DC 20233-0001

Final
DSMD
Signed

December 01, 2005

MEMORANDUM FOR Walter C. Odom, Jr.
Chief, Administrative and Customer Services Division

From: Alan R. Tupek *(Signed)*
Chief, Demographic Statistical Methods Division

Subject: SIPP 2004: Source and Accuracy Statement for 2004 Panel Wave 1 Preliminary (core) Public Use Files (S&A-4)

The attached document is the Source and Accuracy Statement for the 2004 Panel Wave 1 Preliminary (core) Public Use Files.

Attachment

cc:

A. Shields	(ACSD)	C. Nelson	(HHES)
T. Blatt	(DSD)	C. Popoff	
N. McKee		T. Palumbo	
Z. McBride		J. Hisnanick	
J. Eargle		A. Gottschalck	
P. Benton		Q. Wang	
R. Kominski	(POP)	A. Jones Jr.	
M. O'Connell		J. Day	
K. Bauman		J. Tin	
P. Salopek		S. Stern	
J. Fields		M. Weismantle	
		B. Downs	
		C. Gunlicks	(DSMD)
		SIPPB	

DSMD: C:\Documents and Settings\thrif300\Local Settings\Temp\notesE1EF34\2004s&a.wpd

SOURCE AND ACCURACY STATEMENT FOR THE 2004 PRELIMINARY WAVE 1 PUBLIC USE (CORE) FILES FROM THE SURVEY OF INCOME AND PROGRAM PARTICIPATION[1]

SOURCE OF DATA

The data were collected in the 2004 panel of the Survey of Income and Program Participation (SIPP). The population represented (the population universe) in the 2004 SIPP is the civilian noninstitutionalized population living in the United States. The institutionalized population, which is excluded from the population universe, is composed primarily of the population in correctional institutions and nursing homes (91 percent of the 4.1 million institutionalized people in Census 2000). The population includes persons living in group quarters, such as dormitories, rooming houses, and religious group dwellings. Crew members of merchant vessels, Armed Forces personnel living in military barracks, and institutionalized persons, such as correctional facility inmates and nursing home residents, were not eligible to be in the survey. Also, United States citizens residing abroad were not eligible to be in the survey. Foreign visitors who work or attend school in this country and their families were eligible; all others were not eligible to be in the survey. With the exceptions noted above, persons who were at least 15 years of age at the time of the interview were eligible to be in the survey.

The 2004 Panel of the SIPP sample is located in 123 self-representing (SR) primary sampling units (PSU) and 228 non-self-representing (NSR) PSUs. SR PSUs have a probability of selection of one, NSR PSUs have a probability of selection of less than one. Each PSU consists of a county or a small cluster of contiguous counties. Within these PSUs, housing units (HUs) were systematically selected from the master address file (MAF) used for the 2000 decennial census. To account for HUs built within each of the sample areas after the 2000 census, a sample containing clusters of four HUs was drawn of permits issued for construction of residential HUs up until shortly before the beginning of the panel.

In jurisdictions that do not issue building permits or have incomplete addresses, we systematically sampled expected clusters of four HUs which were then listed by field personnel.

Sample households within a given panel are divided into four random subsamples of nearly equal size. These subsamples are called rotation groups and one rotation group is interviewed each month. Each household in the sample was scheduled to be interviewed at 4 month intervals over a period of roughly 5 years beginning in February 2004. The reference period for the questions is the 4-month period preceding the interview month. The most recent month is designated reference 4, the earliest month is reference month 1. In general, one cycle of four interviews covering the entire sample, using the same questionnaire, is called a wave. For example, Wave 1 rotation group 1 of the 2004 Panel was interviewed in February 2004 and data for the reference months October 2003 through January 2004 were collected.

---

[1] For questions or further assistance with the information provided in this document contact Tracy L. Matttingly of the SIPP branch of the Demographic Statistical Methods Division on (301) 763-6445 or via the e-mail using tracy.l.mattingly@census.gov.

In Wave 1 we fielded a sample of about 62,700 HUs. About 11,300 of these HUs were found to be vacant, demolished, converted to nonresidential use, or otherwise ineligible for the survey. Interviews were obtained at about 43,700 of the eligible HUs. We did not interview approximately 7,600 eligible HUs, because the occupants: (1) refused to be interviewed; (2) could not be found at home; (3) were temporarily absent; or (4) were otherwise unavailable. Thus, occupants of about 85 percent of all eligible HUs participated in the first interview of the panel.

For subsequent interviews, only original sample persons (those in Wave 1 sample households and interviewed in Wave 1) and persons living with them are eligible to be interviewed. We will follow original sample persons if they move to a new address, unless the new address is more than 100 miles from a SIPP sample area. Then, we will attempt telephone interviews.

Estimation. We used several stages of weight adjustments in the estimation procedure to derive the SIPP cross-sectional person level weights. We gave each person a base weight (BW) equal to the inverse of probability of selection of a person's household. We applied a noninterview adjustment factor to account for households which were eligible for the sample but which field representatives could not interview in wave 1 (F_{NI}). We used a Duplication Control Factor (DCF) to adjust for subsampling done in the field when the number of sample units is much larger than expected. The last adjustment is the Second Stage Adjustment Factor (F_{2S}). This adjusts estimates to population controls and equalizes husbands' and wives' weights. The 2004 Panel adjusts weights to both national and state level controls.

The final cross-sectional weight is $FW_c = BW * DCF * F_{N1} * F_{2s}$ for Wave 1. Additional details of the weighting process are in *SIPP 2004+: Cross-Sectional Weighting Specifications for Wave 1.*

Population Controls. This survey's estimation procedure adjusts weighted sample results to agree with independently derived population estimates of the civilian noninstitutional population of the United States. We control to independent population estimates in an attempt to reduce our mean square error by partially correcting for undercoverage. To obtain the national family type controls, we take the CPS weights and do a "March type" family equalization. That is, we assign wives' weights to husbands and then proportionally adjust the weights of persons by month, rotation group, race, sex, age, and by the marital and family status of householders. The national and state level population controls are obtained directly from the Population Division. These are prepared each month to agree with the most current set of population estimates that are released as part of the Census Bureau's population estimates and projections program.

The national controls are distributed by demographic characteristics in two ways:

- Age, Sex, and Race (White alone, Black alone, and all other groups combined)
- Age, Sex, and Hispanic Origin

The state controls are distributed by demographic characteristics in three ways.

- State by Age and Sex
- State by Hispanic origin
- State by Race (Black alone, all other groups combined)

The estimates begin with the latest decennial census as the base and incorporate the latest available information on births and deaths along with the latest estimates of net international migration.

The net international migration component in the population estimates includes a combination of:

- legal migration to the U.S.,
- emigration of foreign born and native people from the U.S.,
- net movement between the U.S. and Puerto Rico,
- estimates of temporary migration, and
- estimates of net residual foreign-born population, which include unauthorized migration.

Because the latest available information on these components lags the survey date, to develop the estimate for the survey date, it is necessary to make short-term projections of these components.

Additional Methodology

Use of Weights. There are three primary weights for the analysis of SIPP data. The person month weight (one for each reference month) is for analyzing data at the person level. Everyone in sample in a given reference month has a person month weight. The person month weight of the household reference person is used to analyze data at the household level (a household may consist of related and unrelated persons). The person month weight of the family reference person is the family weight. Use this weight to analyze family level questions. Weights are also available in the public use files for related subfamilies. Chapter 8 of the SIPP Users' Guide: 2001 provides additional information on how to use these weights.

By selecting the appropriate reference month weight an analyst can obtain the average of an item such as income across several calendar months.

> **Example**: using the proper weights, one can estimate the monthly average number of households in a specified income range over December 2003 to January 2004. To estimate monthly averages of a given measure, e.g., total, mean, over a number of consecutive months, sum the monthly estimates and divide by the number of months.

To form an estimate for a particular month, use the <u>reference month</u> weight for the month of interest, summing over all persons or households with the characteristic of interest whose reference period includes the month of interest.

The core wave file contains no weight for characteristics that involve a persons's or household's status over two or more months (such as, number of households with a 50 percent increase in income between December 2003 and January 2004).

ACCURACY OF ESTIMATES

SIPP estimates are based on a sample, they may differ somewhat from the figures that would have been obtained if a complete census had been taken using the same questionnaire, instructions, and enumerators. There are two types of errors possible in an estimate based on a sample survey: sampling and nonsampling. We are able to provide estimates of the magnitude of SIPP sampling error, but this is not true of nonsampling error.

Nonsampling Error. Nonsampling errors can be attributed to many sources:

- inability to obtain information about all cases in the sample;
- definitional difficulties;
- differences in the interpretation of questions;
- inability or unwillingness on the part of the respondents to provide correct information;
- errors made in the following: collection such as in recording or coding the data, processing the data, estimating values for missing data;
- biases resulting from the differing recall periods caused by the interviewing pattern used; and
- undercoverage.

Quality control and edit procedures are used to reduce errors made by respondents, coders and interviewers. More detailed discussions of the existence and control of nonsampling errors in the SIPP can be found in the *SIPP Quality Profile, 1998 SIPP Working Paper Number 230,* issued May 1999.

Undercoverage in SIPP results from missed HUs and missed persons within sample households. It is known that undercoverage varies with age, race, and sex. Generally, undercoverage is larger for males than for females and larger for Blacks than for non-Blacks. Ratio estimation to independent age-race-sex population controls partially corrects for the bias due to survey undercoverage. The independent population controls have been adjusted for undercoverage in the Census. However, biases exist in the estimates to the extent that persons in missed households or missed persons in interviewed households have characteristics different from those of interviewed persons in the same age-race-sex group.

A common measure of survey coverage is the coverage ratio, the estimated population before ratio adjustment divided by the independent population control. Table 1 below shows SIPP coverage ratios for age-sex-race groups for one month, January 2004, prior to the ratio adjustment. The SIPP coverage ratios exhibit some variability from month to month.

Table 1. SIPP Average Coverage Ratios for January 2004 for Age by Race and Sex

	White Only		Black Only		Residual	
Age	Males	Females	Males	Females	Males	Females
0-4	0.87	0.89	0.85	0.76	1.20	1.09
5-9	0.93	0.93	0.84	0.80	1.15	1.06
10-14	0.91	0.91	0.85	0.88	1.11	1.05
15-24	0.82	0.82	0.76	0.79	0.95	0.96
25-34	0.82	0.89	0.75	0.80	0.88	0.95
35-44	0.89	0.90	0.82	0.86	0.99	1.01
45-54	0.89	0.92	0.80	0.91	1.02	1.06
55-64	0.90	0.95	0.86	0.94	1.02	1.15
65+	0.96	0.93	0.94	1.03	0.97	0.98

Comparability with Other Estimates. Caution should be exercised when comparing data from this with data from other SIPP products or with data from other surveys. The comparability problems are caused by such sources as the seasonal patterns for many characteristics, different nonsampling errors, and different concepts and procedures. Refer to the *SIPP Quality Profile* for known differences with data from other sources and further discussions.

Sampling Variability. Standard errors indicate the magnitude of the sampling error. They also partially measure the effect of some nonsampling errors in response and enumeration, but do not measure any systematic biases in the data. The standard errors for the most part measure the variations that occurred by chance because a sample rather than the entire population was surveyed.

USES AND COMPUTATION OF STANDARD ERRORS

Confidence Intervals. The sample estimate and its standard error enable one to construct confidence intervals, ranges that would include the average result of all possible samples with a known probability. For example, if all possible samples were selected, each of these being surveyed under essentially the same conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then:

1. Approximately 68 percent of the intervals from one standard error below the estimate to one standard error above the estimate would include the average result of all possible samples.

2. Approximately 90 percent of the intervals from 1.645 standard errors below the estimate to 1.645 standard errors above the estimate would include the average result of all possible samples.

3. Approximately 95 percent of the intervals from two standard errors below the estimate to two standard errors above the estimate would include the average result of all possible samples.

The average estimate derived from all possible samples is or is not contained in any particular computed interval. However, for a particular sample, one can say with a specified confidence that the average estimate derived from all possible samples is included in the confidence interval.

Hypothesis Testing. Standard errors may also be used for hypothesis testing, a procedure for distinguishing between population characteristics using sample estimates. The most common types of hypotheses tested are 1) the population characteristics are identical versus 2) they are different. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the characteristics are different when, in fact, they are identical.

To perform the most common test, compute the difference $X_A - X_B$, where X_A and X_B are sample estimates of the characteristics of interest. A later section explains how to derive an estimate of the standard error of the difference $X_A - X_B$. Let that standard error be S_{DIFF}. If $X_A - X_B$ is between -1.645 times S_{DIFF} and +1.645 times S_{DIFF}, no conclusion about the characteristics is justified at the 10 percent significance level. If, on the other hand, $X_A - X_B$ is smaller than -1.645 times S_{DIFF} or larger than +1.645 times S_{DIFF}, the observed difference is significant at the 10 percent level. In this event, it is commonly accepted practice to say that the characteristics are different. Of course, sometimes this conclusion will be wrong. When the characteristics are the same, there is a 10 percent chance of concluding that they are different.

Note that as more tests are performed, more erroneous significant differences will occur. For example, at the 10 percent significance level, if 100 independent hypothesis tests are performed in which there are no real differences, it is likely that about 10 erroneous differences will occur. Therefore, the significance of any single test should be interpreted cautiously. A Bonferroni correction can be done to account for this potential problem that consists of dividing your stated level of significance by the number of tests you are performing. This correction results in a conservative test of significance.

Note Concerning Small Estimates and Small Differences. Because of the large standard errors involved, there is little chance that estimates will reveal useful information when computed on a base smaller than 200,000. Care must be taken in the interpretation of small differences since even a small amount of nonsampling error can cause a borderline difference to appear significant or not, thus distorting a seemingly valid hypothesis test.

Calculating Standard Errors for SIPP Estimates. There are three main ways we calculate the Standard Errors (SEs) for SIPP Estimates. They are as follows:

- Direct estimates using replicate weighting methods;
- Generalized variance function parameters (denoted as *a* and *b*); and
- Simplified tables of SEs based on the *a* and *b* parameters.

While the replicate weight methods provide the most accurate variance estimates, this approach requires more computing resources and more expertise on the part of the user. The Generalized Variance Function (GVF) parameters provide a method of balancing accuracy with resource usage as well as smoothing effect on SE estimates across time. SIPP uses the Replicate Weighting Method to produce GVF parameters(see K. Wolter, *Introducation to Variance Estimation*, Chapter 5 for more information). The GVF parameters are used to create the simplified tables of SEs.

Standard Error Parameters and Tables and Their Use. Most SIPP estimates have greater standard errors than those obtained through a simple random sample because of its two-stage cluster sample design. To derive standard errors that would be applicable to a wide variety of estimates and could be prepared at a moderate cost, a number of approximations were required.

These approximations are created by modeling predicted variances from replicate based direct estimates of groups of items of interest to analysts (domains). The estimates are grouped by common subject matter (e.g., estimates of poverty and program participation). Within each of these domains estimates with similar variance characteristics are used to create the model based approximations (a and b parameters). These *a* and *b* parameters vary according to wave and characteristic as well as the demographic subgroup of the group to which the estimate applies. Because the actual standard error behavior was not identical for all characteristics and groups, the standard errors computed using these parameters provide an indication of the order of magnitude of the standard error estimate for a specific group. Table 2 provides tables of base *a* and *b* parameters by domain to be used for the 2004 panel Wave 1 preliminary file estimates.

Adjusting for Calendar Months with Less than Four Rotations. When estimates for months with less than four rotations worth of data are constructed from a wave file, factors greater than 1 must be applied. Multiply the sum by a factor to account for the number of rotations contributing data for the month. This factor equals four divided by the number of rotations contributing data for the month. For example, December 2003 data are only available from rotations 1-3 for Wave 1 of the 2004 Panel, so a factor of 4/3 must be applied. A list of appropriate factors are in Table 3.

> **Example.** Use Table 3 (if needed) to select the adjustment factor appropriate to the wave. Multiply this factor by the *a* and *b* base parameters of Table 3 to produce *a* and *b* parameters for the variance estimate for a specific subgroup and reference period. For example, for Wave 1 of the 2004 panel the base *a* and *b* parameters for total number of households are -0.0000278785 and 3,129, respectively. Using Table 3 for Wave 1, the factor for November 2003 is 2 *since only 2 rotation months of data are available*. So the *a* and *b* parameters for the variance estimate of a white household characteristic in November 2003 based on Wave 1 are:

$$-0.0000278785 \times 2 = -0.000055757 \text{ and } 3{,}129 \times 2 = 6{,}258, \text{ respectively.}$$

Similarly, the factor for the last quarter of 2003 is 1.8519 (Table 3) since the only data available are the 6 rotation months from Wave 1 (rotation 1 provides three rotation months, rotation 2 provides two rotation months, and rotation 3 provides one rotation month of data.) So the *a* and *b* parameters for the variance estimate of a white household characteristic in the last quarter of 2003 are are:

$$-0.0000278785 \times 1.8519 = -0.0000516282 \text{ and } 3{,}129 \times 1.8519 = 5{,}794, \text{ respectively.}$$

The *a* and *b* parameters may be used to calculate the standard error for estimated numbers and percentages. Because the actual standard error behavior was not identical for all estimates within a group, the standard errors computed from these parameters provide an indication of the order of magnitude of the standard error for any specific estimate. Methods for using these parameters for computation of approximate standard errors are given in the following sections.

For those users who wish further simplification, we have also provided base standard errors for estimates of totals and percentages in Tables 4 through 7. Note that these base standard errors only apply when data from all four rotations are used and must be adjusted by an f factor provided in Table 2. The standard errors resulting from this simplified approach are less accurate. Methods for using these parameters and tables for computation of standard errors are given in the following sections.

The procedures described below apply only to reference month estimates or averages of reference month estimates. Refer to the sections "Use of Weights" and "Adjusting for Calendar Months with Less than Four Rotations" for a more detailed discussion of the construction of estimates.

Variance stratum codes and half sample codes are included in the data sets to enable the user to compute the variances directly and more accurately by other methods. William G. Cochran provides a list of references discussing the application of these variables. (See Sampling Techniques, 3rd Ed., New York: John Wiley and Sons, 1977, p. 321.)

Standard Errors of Estimated Numbers. The approximate standard error, s_x, of an estimated number of persons, households, families, unrelated individuals and so forth, can be obtained in two ways. Both apply when data from all four rotations are used to make the estimate. However, only the second method (formula 2) should be used when less than four rotations of data are available for the estimate. Note that neither method should be applied to dollar values.

The standard error may be obtained by the use of the formula

$$s_x = fs \qquad (1)$$

where f is the appropriate f factor from Table 2, and s is the base standard error on the estimate obtained by interpolation from Table 4 or 5. Alternatively, s_x may be approximated by the formula:

$$s_x = \sqrt{ax^2 + bx} \qquad (2)$$

This formula was used to calculate the base standard errors in Tables 6 and 7. Here x is the size of the estimate and a and b are the parameters from Table 2 which are associated with the characteristic being estimated (and the wave which applies). Use of formula 2 will generally provide more accurate results than the use of formula 1.

Illustration.

Suppose SIPP estimates based on Wave 1 of the 2004 panel show that there were 1,700,000 black households with monthly household income above $4,000 in January 2004. The appropriate parameters and factor from Table 2 and the appropriate general standard error from Table 4 are:

$$a = -0.0002273351 \qquad b = 3{,}129 \qquad f = 0.843 \qquad s = 66{,}640$$

Using formula 1, the approximate standard error is:

$$s_x = (0.843)(66{,}640) = 56{,}177$$

Using formula 2, the approximate standard error is:

$$\sqrt{(-0.0002273351)(1{,}700{,}000)^2 + (3{,}129)(1{,}700{,}000)} = 68{,}281$$

Using the standard error based on formula 2, the approximate 90-percent confidence interval as shown by the data is from 1,631,719 to 1,768,281. Therefore, a conclusion that the average estimate derived from all possible samples lies within a range computed in this way would be correct for roughly 90% of all samples.

Standard Error of a Mean. A mean is defined here to be the average quantity of some item (other than persons, families, or households) per person, family or household. For example, it could be the average monthly household income of females age 25 to 34. The standard error of a mean can be approximated by formula 3 below. Because of the approximations used in developing formula 3, an estimate of the standard error of the mean obtained from this formula will generally underestimate the true standard error. The formula used to estimate the standard error of a mean $\overline{x}$ is:

$$s_{\overline{x}} = \sqrt{\left(\frac{b}{y}\right) s^2} \qquad (3)$$

where y is the size of the base, s^2 is the estimated population variance of the item and b is the parameter associated with the particular type of item.

The population variance s^2 may be estimated by one of two methods. In both methods, we assume x_i is the value of the item for unit "i." (Unit may be person, family, or household). To use the first method, the range of values for the item is divided into "c" intervals. The upper and lower boundaries of interval j are Z_{j-1} and Z_j, respectively. Each unit is placed into one of "c" groups such that $Z_{j-1} < x_i \le Z_j$.

The estimated population variance, s^2, is given by the formula:

$$s^2 = \sum_{j=1}^{c} p_j m_j^2 - \overline{x}^2, \tag{4}$$

where p_j is the estimated proportion of units in group j, and $m_j = (Z_{j-1} + Z_j) / 2$. The most representative value of the item in group j is assumed to be m_j. If group "c" is open-ended, or there is no upper interval boundary exists, then an approximate value for m_c is

$$m_c = \frac{3}{2} Z_{c-1}.$$

The mean, $\overline{x}$ can be obtained using the following formula:

$$\overline{x} = \sum_{j=1}^{c} p_j m_j$$

In the second method, the estimated population mean, $\overline{x}$, and variance, s^2 are given by:

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

$$s^2 = \frac{\sum_{i=1}^{n} w_i x_i^2}{\sum_{i=1}^{n} w_i} - \overline{x}^2, \tag{5}$$

where there are n units with the item of interest and w_i is the final weight for unit "I". (Note that $\sum w_i = y$ in formula 3.)

Illustration.

Applying formula 4 and the mean monthly cash income of $2,530 for persons aged 25 to 34 from the example data in Table 8, the approximate population variance, s^2, is:

$$s^2 = \left(\frac{1{,}371}{39{,}851}\right)(150)^2 + \left(\frac{1{,}651}{39{,}851}\right)(450)^2 + \ldots\ldots +$$

$$\left(\frac{1{,}493}{39{,}851}\right)(9{,}000)^2 - (2{,}530)^2 = 3{,}159{,}887.$$

Using formula 3 and a base *b* parameter of 4,263, the estimated standard error of a mean $\bar{x}$ is:

$$s_{\bar{x}} = \sqrt{\left(\frac{4{,}263}{39{,}851{,}000}\right)(3{,}159{,}887)} = \$18.39$$

Standard error of an aggregate. An aggregate is defined to be the total quantity of an item summed over all the units in a group. The standard error of an aggregate can be approximated using formula 6.

As with the estimate of the standard error of a mean, the estimate of the standard error of an aggregate will generally underestimate the true standard error. Let y be the size of the base, s^2 be the estimated population variance of the item obtained using formula (4) or (5) and b be the parameter associated with the particular type of item. The standard error of an aggregate is:

$$s_x = \sqrt{(b)\ (y)s^2} \tag{6}$$

Standard Errors of Estimated Percentages. The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends upon both the size of the percentage and the size of the total upon which the percentage is based. Estimated percentages are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are 50 percent or more, e.g., the percent of people employed is more reliable than the estimated number of people employed. When the numerator and denominator of the percentage have different parameters, use the parameter (and appropriate factor) of the numerator. If proportions are presented instead of percentages, note that the standard error of a proportion is equal to the standard error of the corresponding percentage divided by 100.

There are two types of percentages commonly estimated. The first is the percentage of persons, families or households sharing a particular characteristic such as the percent of persons owning their own home. The second type is the percentage of money or some similar concept held by a particular group of persons or held in a particular form. Examples are the percent of total wealth held by persons with high income and the percent of total income received by persons on welfare.

For the percentage of persons, families, or households, the approximate standard error, $s_{(x,p)}$, of the estimated percentage p can be obtained by the formula:

$$s_{(x,p)} = fs \tag{7}$$

when data from all four rotations are used to estimate p.

In this formula, f is the appropriate f factor from Table 2 (for the appropriate wave) and s is the base standard error of the estimate from Table 6 or 7.

Alternatively, it may be approximated by the formula:

$$s_{(x,p)} = \sqrt{\frac{b}{x}\,(p)\,(100-p)} \tag{8}$$

from which the standard errors in Tables 6 and 7 were calculated. Here x is the size of the subclass of social units which is the base of the percentage, p is the percentage ($0 < p < 100$), and b is the parameter associated with the characteristic in the numerator. Use of this formula will give more accurate results than use of formula 7 above and should be used when data from less than four rotations are used to estimate p.

<u>Illustration.</u>

Suppose that, 6.7 percent of the 16,812,000 persons in nonfarm households with a mean monthly household cash income of $4,000 to $4,999, were black. Using formula 8 and a b parameter of 4,475 and a factor of 1 from Table 3, the approximate standard error is:

$$\sqrt{(4{,}475/16{,}812{,}000)\ (6.7)\ (100-6.7)} = 0.41 \text{ percent}$$

Consequently, the 90 percent confidence interval as shown by these data is from 6.03 to 7.37 percent.

For percentages of money, a more complicated formula is required. A percentage of money will usually be estimated in one of two ways. It may be the ratio of two aggregates:

$$p_I = 100\ (X_A\ /\ X_N)$$

or it may be the ratio of two means with an adjustment for different bases:

$$p_I = 100\ (\hat{p}_A\ \bar{X}_A\ /\ \bar{X}_N)$$

where x_A and x_N are aggregate money figures, $\overline{x}_A$ and $\overline{x}_N$ are mean money figures, and $\hat{p}_A$ is the estimated number in group A divided by the estimated number in group N. In either case, we estimate the standard error as

$$s_I = \sqrt{\left(\frac{\hat{p}_A \overline{x}_A}{\overline{x}_N}\right)^2 \left[\left(\frac{s_p}{\hat{p}_A}\right)^2 + \left(\frac{s_A}{\overline{x}_A}\right)^2 + \left(\frac{s_B}{\overline{x}_N}\right)^2\right]} \, , \tag{9}$$

where s_p is the standard error of $\hat{p}_A$, s_A is the standard error of $\overline{x}_A$ and s_B is the standard error of $\overline{x}_N$. To calculate s_p, use formula 8. The standard errors of $\overline{x}_N$ and $\overline{x}_A$ may be calculated using formula 3.

It should be noted that there is frequently some correlation between $\hat{p}_A$, $\overline{x}_N$, and $\overline{x}_A$. Depending on the magnitude and sign of the correlations, the standard error will be over or underestimated.

<u>Illustration.</u>

Suppose that 9.8% of the households own rental property, the mean value of rental property is \$72,121, the mean value of assets is \$78,734, and the corresponding standard errors are 0.19 %, \$5,799, and \$2,867, respectively. In total there are 86,790,000 households. Then, the percent of all household assets held in rental property is:

$$= 100 \left((0.098) \frac{72121}{78734} \right) = 9.0\%$$

Using formula (9), the appropriate standard error is:

$$s_I = \sqrt{\left(\frac{(0.098)(72121)}{78734}\right)^2 \left[\left(\frac{0.0019}{0.098}\right)^2 + \left(\frac{5799}{72121}\right)^2 + \left(\frac{2867}{78734}\right)^2\right]}$$

$= 0.008 \quad = 0.8\%$

Standard Error of a Difference. The standard error of a difference between two sample estimates is approximately equal to

$$s_{(x-y)} = \sqrt{s_x^2 + s_y^2} \tag{10}$$

where s_x and s_y are the standard errors of the estimates x and y. The estimates can be numbers, percents, ratios, etc. The above formula assumes that the correlation coefficient between the characteristics estimated by x and y is zero. If the correlation is really positive (negative), then this assumption will tend to cause overestimates (underestimates) of the true standard error.

Illustration.

Suppose that SIPP estimates show the number of persons age 35-44 years with monthly cash income of $4,000 to $4,999 was 3,186,000 and the number of persons age 25-34 years with monthly cash income of $4,000 to $4,999 in the same time period was 2,619,000. The standard errors of these numbers are approximately 115,689 and 105,029, respectively. The difference in sample estimates is 567,000 and using formula 10, the approximate standard error of the difference is:

$$\sqrt{(115{,}689)^2 + (105{,}029)^2} = 156{,}253$$

It is desired to test at the 10 percent significance level whether the number of persons with monthly cash income of $4,000 to $4,999 were different for persons age 35-44 years than for persons age 25-34 years. To perform the test, compare the difference of 567,000 to the product $1.645 \times 156{,}253 = 257{,}036$. Since the difference is greater than 1.645 times the standard error of the difference, the data show that the two age groups are significantly different at the 10 percent significance level.

Standard Error of a Median. The median quantity of some item such as income for a given group of persons, families, or households is that quantity such that at least half the group have as much or more and at least half the group have as much or less. The sampling variability of an estimated median depends upon the form of the distribution of the item as well as the size of the group. To calculate standard errors on medians, the procedure described below may be used.

An approximate method for measuring the reliability of an estimated median is to determine a confidence interval about it. (See the section on sampling variability for a general discussion of confidence intervals.) The following procedure may be used to estimate the 68-percent confidence limits and hence the standard error of a median based on sample data.

1. Determine, using either formula 7 or formula 8, the standard error of an estimate of 50 percent of the group.

2. Add to and subtract from 50 percent the standard error determined in step 1.

3. Using the distribution of the item within the group, calculate the quantity of the item such that the percent of the group with more of the item is equal to the smaller percentage found in step 2. This quantity will be the upper limit for the 68-percent confidence interval. In a similar fashion, calculate the quantity of the item such that the percent of the group with more of the item is equal to the larger percentage found in step 2. This quantity will be the lower limit for the 68-percent confidence interval.

4. Divide the difference between the two quantities determined in step 3 by two to obtain the standard error of the median.

To perform step 3, it will be necessary to interpolate. Different methods of interpolation may be used. The most common are simple linear interpolation and Pareto interpolation. The appropriateness of the method depends on the form of the distribution around the median. If density is declining in the area, then we recommend Pareto interpolation. If density is fairly constant in the area, then we recommend linear interpolation. Note, however, that Pareto interpolation can never be used if the interval contains zero or negative measures of the item of interest. Interpolation is used as follows. The quantity of the item such that *p* percent have more of the item is:

$$X_{pN} = \exp\left[\left(\mathrm{Ln}\left(\frac{pN}{N_1}\right) \Big/ \mathrm{Ln}\left(\frac{N_2}{N_1}\right)\right) \mathrm{Ln}\left(\frac{A_2}{A_1}\right)\right] A_1 \qquad (11)$$

if Pareto Interpolation is indicated and:

$$X_{pN} = \left[\frac{PN - N_1}{N_2 - N_1} \quad (A_2 - A_1) \ + A_1\right] \qquad (12)$$

if linear interpolation is indicated, where:

N	is the size of the group,
A_1 and A_2	are the lower and upper bounds, respectively, of the interval in which X_{pN} falls
N_1 and N_2	are the estimated number of group members owning more than A_1 and A_2, respectively
exp	refers to the exponential function and
Ln	refers to the natural logarithm function

Illustration.

To illustrate the calculations for the sampling error on a median, return to Table 8. The median monthly income for this group is \$2,158. The size of the group is 39,851,000.

1. Using formula 8 (with b = 4,263), the standard error of 50 percent on a base of 39,851,000 is about 0.5 percentage points.

2. Following step 2, the two percentages of interest are 49.5 and 50.5.

3. By examining Table 8, we see that the percentage 49.5 falls in the income interval from \$2,000 to \$2,499. (Since 55.5% receive more than \$2,000 per month, the dollar value corresponding to 49.5 must be between \$2,000 and \$2,500). Thus, A_1 = \$2,000, A_2 = \$2,500, N_1 = 22,106,000, and N_2 = 16,307,000.

In this case, we decided to use Pareto interpolation. Therefore, the upper bound of a 68% confidence interval for the median is

$$\$2{,}000 \exp\left[\left(\mathrm{Ln}\left(\frac{(.495)(39{,}851{,}000)}{22{,}106{,}000}\right) \Big/ \mathrm{Ln}\left(\frac{16{,}307{,}000}{22{,}106{,}000}\right)\right)\mathrm{Ln}\left(\frac{2{,}500}{2{,}000}\right)\right] = \$2{,}174$$

Also by examining Table 8, we see that 50.5 falls in the same income interval. Thus, A_1, A_2, N_1 and N_2 are the same. We also use Pareto interpolation for this case. So the lower bound of a 68% confidence interval for the median is

$$\$2{,}000 \exp\left[\left(\mathrm{Ln}\left(\frac{(.505)(39{,}851{,}000)}{22{,}106{,}000}\right) \Big/ \mathrm{Ln}\left(\frac{16{,}307{,}000}{22{,}106{,}000}\right)\right)\mathrm{Ln}\left(\frac{2{,}500}{2{,}000}\right)\right] = \$2{,}142$$

Thus, the 68-percent confidence interval on the estimated median is from \$2,142 to \$2,174. An approximate standard error is

$$\frac{\$2{,}174 - \$2{,}142}{2} = \$16$$

Standard Errors of Ratios of Means and Medians. The standard error for a ratio of means or medians is approximated by:

$$s_{\frac{x}{y}} = \sqrt{\left(\frac{x}{y}\right)^2 \left[\left(\frac{s_y}{y}\right)^2 + \left(\frac{s_x}{x}\right)^2\right]} \tag{13}$$

where x and y are the means or medians, and s_x and s_y are their associated standard errors. Formula 13 assumes that the means are not correlated. If the correlation between the population means estimated by x and y are actually positive (negative), then this procedure will tend to produce overestimates (underestimates) of the true standard error for the ratio of means.

Standard Errors Using SAS or SPSS. Standard errors and their associated variance, calculated by SAS or SPSS statistical software package, do not accurately reflect the SIPP's complex sample design. Erroneous conclusions will result if these standard errors are used directly. We provide adjustment factors by characteristics that should be used to correctly compensate for likely under-estimates. The factors called DEFF available in Table 2, must be applied to SAS or SPSS generated variances. The square root of DEFF can be directly applied to similarly generated standard errors. These factors approximate design effects which adjust statistical measures for sample designs more complex than simple random sample.

Table 2. GVF Parameters for the 2004 Panel Wave 1 Preliminary File

Domain	a parameter	b parameter	DEFF	f
Poverty and Program Participation, Persons 15+				
Total	-0.0000147427	3336	2.03	0.979
Male	-0.0000305564	3336		
Female	-0.0000284871	3336		
Income and Labor Force Participation, Persons 15+				
Total	-0.0000151714	3433	2.09	0.993
Male	-0.0000314448	3433		
Female	-0.0000293154	3433		
Other Persons 0+				
Total (or White)	-0.0000121186	3478	2.12	1.000
Male	-0.0000248014	3478		
Female	-0.0000236980	3478		
Black Persons 0+	-0.0000867450	3118	1.89	0.896
Male	-0.0001870868	3118		
Female	-0.0001617356	3118		
Hispanic Persons 0+	-0.000109909	4407	2.68	1.267
Male	-0.000214642	4407		
Female	-0.000225251	4407		
Households				
Total (or White)	-0.0000278785	3129	1.904	1.000
Black	-0.0002273351	3129		
Hispanic	-0.0002673705	3129		

Notes on Domain Usage:

Poverty and Program Participation	Use these parameters for estimates concerning poverty rates, welfare program participation (e.g., foodstamp, SSI, TANF), and other programs for adults with low incomes.
Income and Labor Force	These parameters are for estimates concerning income, sources of income, labor force participation, economic well being other than poverty, employment related estimates (e.g., occupation, hours worked a week), and other income, job, or employment related estimates.
Other Persons	Use the “Other Persons” parameters for estimates of total (or white) persons aged 0+ in the labor force, and all other characteristics not specified in this table, for the total or white population.
Black/Hispanic Persons	Use these parameters for estimates of Black and Hispanic persons 0+.
Households	Use these parameters for all household level estimates.

Table 3. Factors to be Applied to Table 3 Base Parameters to Obtain Parameters for Various Reference Periods

Number of Available Rotation Months[2]	Factor
Monthly Estimate	
1	4.0000
2	2.0000
3	1.3333
4	1.0000
Quarterly Estimate	
6	1.8519
8	1.4074
9	1.2222
10	1.0494
11	1.0370
12	1.0000

[2] The number of available rotation months for a given estimate is the sum of the number of rotations available for each month of the estimates.

Table 4. Base Standard Errors of Estimated Numbers of Household or Families.

Size of Estimate	Standard Error	Size of Estimate	Standard Error
200,000	24,994	30,000,000	262,258
300,000	30,597	40,000,000	283,821
500,000	39,466	50,000,000	294,540
750,000	48,281	60,000,000	295,597
1,000,000	55,688	70,000,000	287,098
2,000,000	78,400	80,000,000	268,137
3,000,000	95,583	90,000,000	236,208
5,000,000	122,262	95,000,000	213,662
7,500,000	147,984	99,500,000	187,966
10,000,000	168,826	105,000,000	145,549
15,000,000	201,649	110,000,000	82,826
25,000,000	246,578	112,236,860	726

Note: These estimates are calculations using the Households Total (or White) a and b parameters from Table 2.

Table 5. Base Standard Errors of Estimated Numbers of Persons.

Size of Estimate	Standard Error	Size of Estimate	Standard Error
200,000	26,365	110,000,000	485,742
300,000	32,285	120,000,000	492,801
500,000	41,665	130,000,000	497,329
750,000	51,007	140,000,000	499,396
1,000,000	58,872	150,000,000	499,031
2,000,000	83,112	160,000,000	496,230
3,000,000	101,612	170,000,000	490,951
5,000,000	130,717	180,000,000	483,113
7,500,000	159,384	190,000,000	472,588
10,000,000	183,216	200,000,000	459,192
15,000,000	222,359	210,000,000	442,664
25,000,000	281,737	220,000,000	422,636
30,000,000	305,669	230,000,000	398,582
40,000,000	346,021	240,000,000	369,717
50,000,000	378,951	250,000,000	334,797
60,000,000	406,267	260,000,000	291,658
70,000,000	429,044	264,000,000	271,249
80,000,000	447,974	270,000,000	235,831
90,000,000	463,529	280,000,000	154,091
100,000,000	476,040	286,997,543	0

Note: These estimates are calculations using the Other Persons 0+ a and b parameters from Table 2.

To calculate the standard for another domain multiply the standard error from this table by the appropriate f factor from Table 2.

Table 6. Base Standard Errors for Percentages of Households or Families

	Estimated Percentages					
Base of Estimated Percentages	≤1 or ≥99	2 or 98	5 or 95	10 or 90	25 or 75	50
200,000	1.24%	1.75%	2.73%	3.75%	5.42%	6.25%
300,000	1.02%	1.43%	2.23%	3.06%	4.42%	5.11%
500,000	0.79%	1.11%	1.72%	2.37%	3.43%	3.96%
750,000	0.64%	0.90%	1.41%	1.94%	2.80%	3.23%
1,000,000	0.56%	0.78%	1.22%	1.68%	2.42%	2.80%
2,000,000	0.39%	0.55%	0.86%	1.19%	1.71%	1.98%
3,000,000	0.32%	0.45%	0.70%	0.97%	1.40%	1.61%
5,000,000	0.25%	0.35%	0.55%	0.75%	1.08%	1.25%
7,500,000	0.20%	0.28%	0.45%	0.61%	0.88%	1.02%
10,000,000	0.17%	0.25%	0.39%	0.53%	0.77%	0.88%
15,000,000	0.14%	0.20%	0.31%	0.43%	0.63%	0.72%
25,000,000	0.11%	0.16%	0.24%	0.34%	0.48%	0.56%
30,000,000	0.10%	0.14%	0.22%	0.31%	0.44%	0.51%
40,000,000	0.08%	0.12%	0.19%	0.27%	0.38%	0.44%
50,000,000	0.07%	0.11%	0.17%	0.24%	0.34%	0.40%
60,000,000	0.07%	0.10%	0.16%	0.22%	0.31%	0.36%
70,000,000	0.06%	0.09%	0.15%	0.20%	0.29%	0.33%
80,000,000	0.06%	0.09%	0.14%	0.19%	0.27%	0.31%
90,000,000	0.05%	0.08%	0.13%	0.18%	0.26%	0.29%
105,000,000	0.05%	0.08%	0.12%	0.16%	0.24%	0.27%
110,000,000	0.05%	0.07%	0.12%	0.16%	0.23%	0.27%
112,236,860	0.05%	0.07%	0.12%	0.16%	0.23%	0.26%

Note: These estimates are calculations using the Households Total (or White) b parameter from Table 2.

Table 7. Base Standard Errors for Percentages of Persons.

Base of Estimated Percentages	Estimated Percentages ≤1 or ≥99	2 or 98	5 or 95	10 or 90	25 or 75	50
200,000	1.31%	1.85%	2.87%	3.96%	5.71%	6.59%
300,000	1.07%	1.51%	2.35%	3.23%	4.66%	5.38%
500,000	0.83%	1.17%	1.83%	2.50%	3.61%	4.17%
750,000	0.67%	0.95%	1.48%	2.04%	2.95%	3.40%
1,000,000	0.59%	0.83%	1.29%	1.77%	2.55%	2.95%
2,000,000	0.42%	0.58%	0.91%	1.25%	1.81%	2.09%
3,000,000	0.34%	0.48%	0.74%	1.02%	1.47%	1.70%
5,000,000	0.26%	0.37%	0.57%	0.79%	1.14%	1.32%
7,500,000	0.21%	0.30%	0.47%	0.65%	0.93%	1.08%
10,000,000	0.19%	0.26%	0.41%	0.56%	0.81%	0.93%
15,000,000	0.15%	0.21%	0.33%	0.46%	0.66%	0.76%
25,000,000	0.12%	0.17%	0.26%	0.35%	0.51%	0.59%
30,000,000	0.11%	0.15%	0.23%	0.32%	0.47%	0.54%
40,000,000	0.09%	0.13%	0.20%	0.28%	0.40%	0.47%
50,000,000	0.08%	0.12%	0.18%	0.25%	0.36%	0.42%
60,000,000	0.08%	0.11%	0.17%	0.23%	0.33%	0.38%
70,000,000	0.07%	0.10%	0.15%	0.21%	0.31%	0.35%
100,000,000	0.06%	0.08%	0.13%	0.18%	0.26%	0.29%
110,000,000	0.06%	0.08%	0.12%	0.17%	0.24%	0.28%
120,000,000	0.05%	0.08%	0.12%	0.16%	0.23%	0.27%
130,000,000	0.05%	0.07%	0.11%	0.16%	0.22%	0.26%
140,000,000	0.05%	0.07%	0.11%	0.15%	0.22%	0.25%
150,000,000	0.05%	0.07%	0.10%	0.14%	0.21%	0.24%
160,000,000	0.05%	0.07%	0.10%	0.14%	0.20%	0.23%
170,000,000	0.05%	0.06%	0.10%	0.14%	0.20%	0.23%
180,000,000	0.04%	0.06%	0.10%	0.13%	0.19%	0.22%
190,000,000	0.04%	0.06%	0.09%	0.13%	0.19%	0.21%
200,000,000	0.04%	0.06%	0.09%	0.13%	0.18%	0.21%
210,000,000	0.04%	0.06%	0.09%	0.12%	0.18%	0.20%
220,000,000	0.04%	0.06%	0.09%	0.12%	0.17%	0.20%
230,000,000	0.04%	0.05%	0.08%	0.12%	0.17%	0.19%
240,000,000	0.04%	0.05%	0.08%	0.11%	0.16%	0.19%
250,000,000	0.04%	0.05%	0.08%	0.11%	0.16%	0.19%
280,000,000	0.04%	0.05%	0.08%	0.11%	0.15%	0.18%
286,997,543	0.03%	0.05%	0.08%	0.10%	0.15%	0.17%

Note: These estimates are calculations using the Other Persons 0+ a and b parameter from Table 2.

To calculate the standard for another domain multiply the standard error from this table by the appropriate f factor from Table 2.

Table 8. Distribution of Monthly Cash Income Among People 25 to 34 Years Old (Not Actual Data, Only Use for Calculation Illustrations)

	Interval of Monthly Cash Income												
	Under $300	$300 to $599	$600 to $899	$900 to $1,119	$1,200 to $1,499	$1,500 to $1,999	$2,000 to $2,499	$2,500 to $2,999	$3,000 to $3,499	$3,500 to $3,999	$4,000 to $4,999	$5,000 to $5,999	$6,000 and Over
Number of People in Each Interval (in thousands)	1,371	1,651	2,259	2,734	3,452	6,278	5,799	4,730	3,723	2,519	2,619	1,223	1,493
Cumulative of People with at Least as Much as Lower Bound of Each Interval (in thousands)	39,851 (Total People)	38,480	36,829	34,570	31,836	28,384	22,106	16,307	11,577	7,854	5,335	2,716	1,493
Percent of People with at Least as Much as Lower Bound of Each Interval	100	96.6	92.4	86.7	79.9	71.2	55.5	40.9	29.1	19.7	13.4	6.8	3.7